Document de travail - Modèle processus gaussiens

Arthur Leroy

10 mars, 2020

1 Une cargaison de notations

- M le nombre d'individus,
- N le nombre d'observation par individu,
- i, j les indices pour les individus,
- k, l les indices pour les observations,

On dispose d'un échantillon $\{(y_i^N, t_i^N)\}_{i=1,\dots,M}$, tel que:

- $t^N = (t_1, \dots, t_N)^T$ le vecteur des temps de l'individu i (communs à tous),
- $y_i^N = y_i(t^N) = (y_i(t_1), \dots, y_i(t_N))^T$ le vecteur des observations de l'individu i,
- $t^O \in \mathbb{R}^O, \ t^O \subset t^N$ le vecteur des temps d'observation d'un nouvel individu *,
- $t^P \in \mathbb{R}^P, \ t^P \subset t^N$ le vecteur des temps à prédire d'un nouvel individu *,
- $O \ + P = N$ et $t_i^O \cup t_i^P = t_i^N, \forall i$
- $y_i^O = y_i(t^O), \, y_i^P = y_i(t^P) \ , \forall i$
- K_{θ_0} une noyau d'hyperparamètres θ_0
- $(\Sigma_{\theta_i})_i$ un ensemble de noyau de même forme et d'hyperparamètres respectifs $(\theta_i)_i$
- $\sigma^2 \in R$
- $\Theta = \{\theta_0, (\theta_i)_i, \sigma^2\}$ le vecteur des hyperparamètres du modèle

L'utilisation des indices avec les autres objets est cohérente avec celle ci-dessus.

2 Le modèle et les hypothèses

On pose le modèle suivant:

$$\forall i, \forall t, y_i(t) = \mu_0(t) + f_i(t) + \epsilon_i(t)$$

- $\mu_0(\cdot) \sim GP(0, K_{\theta_0}(\cdot, \cdot))$
- $f_i(\cdot) \sim GP(0, \Sigma_{\theta_i}(\cdot, \cdot)), \ (f_i)_i \perp\!\!\!\perp$
- $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2), \ (\epsilon_i)_i \perp\!\!\!\perp, \forall t \in R$
- $\mu_0 \perp\!\!\!\perp (f_i)_i$

On note: $\Psi_i(\cdot, \cdot) = \Sigma_{\theta_i}(\cdot, \cdot) + \sigma^2 I_d$

$$y_i(\cdot) | \mu_0 \sim GP(\mu_0(\cdot), \Psi_i(\cdot, \cdot)), \ (y_i | \mu_0)_i \perp\!\!\!\perp$$

Si on applique cela à notre échantillon, on a la loi a priori suivante:

$$y_i^N | \mu_0^N \sim \mathcal{N}(\mu_0^N, \Psi_i^N)$$

avec

$$\Psi_i^N = \begin{pmatrix} \Psi_i^{O,O} & \Psi_i^{O,P} \\ \Psi_i^{P,O} & \Psi_i^{P,P} \end{pmatrix}$$

3 Prédiction pour un nouvel individu partiellement observé

3.1 Calcul de la loi a posteriori

L'objectif réside dans la prédiction de $p(y_*^P | y_*^O, (y_i^N)_i)$, la loi des temps à prédire sachant les temps observés du nouvel individu * et les observations des autres individus.

Tout d'abord, on sait que:

$$\begin{aligned} p(y_*^N | (y_i^N)_i) &= \int p(y_*^N, \mu_0^N | (y_i^N)_i) \ d\mu_0^N \\ &= \int p(y_*^N | (y_i^N)_i, \mu_0^N) p(\mu_0^N | (y_i^N)_i) \ d\mu_0^N \\ &\underbrace{=}_{(y_i|\mu_0)_i \perp\!\!\!\perp} \int p(y_*^N | \mu_0^N) p(\mu_0^N | (y_i^N)_i) \ d\mu_0^N \end{aligned}$$

Et par définition du modèle,

$$p(y_*^N | \mu_0^N) = \mathcal{N}(\mu_0^N, \Psi_*^N) = \mathcal{N}\left(\begin{bmatrix} \mu_0^O \\ \mu_0^P \end{bmatrix}, \begin{pmatrix} \Psi_*^{O,O} & \Psi_*^{O,P} \\ \Psi_*^{P,O} & \Psi_*^{P,P} \end{pmatrix} \right)$$

De plus, si on fait l'hypothèse (qui sera démontrée dans le paragraphe 2 sur l'apprentissage) que:

$$p(\mu_0^N | (y_i^N)_i) = \mathcal{N}(m^N, \hat{K}^N) = \mathcal{N}\left(\begin{bmatrix} m^O \\ m^P \end{bmatrix}, \begin{pmatrix} \hat{K}^{O,O} & \hat{K}^{O,P} \\ \hat{K}^{P,O} & \hat{K}^{P,P} \end{pmatrix} \right)$$

Avec:

- $\hat{K}^N = \left((K^N)^{-1} + \sum\limits_{i=1}^{M} (\Psi_i^N)^{-1} \right)^{-1}$
- $m^N = \hat{K}^N \sum\limits_{i=1}^{M} (\Psi_i^N)^{-1} y_i^N$

Alors on obtient que:

$$p(y_*^N | (y_i^N)_i) = \mathcal{N}(m_*^N, \Gamma_*^N) = \mathcal{N}\left(\begin{bmatrix} m_*^O \\ m_*^P \end{bmatrix}, \begin{pmatrix} \Gamma_*^{O,O} & \Gamma_*^{O,P} \\ \Gamma_*^{P,O} & \Gamma_*^{P,P} \end{pmatrix} \right) \tag{1}$$

Avec:

- $m_*^N = m^N$
- $\Gamma_*^N = \Psi_*^N + \hat{K}^N$

De plus, en utilisant la formule de prédiction des processus Gaussiens (voir Bishop section 2.3.1), on obtient que:

$$p(y_*^P | y_*^O, (y_i^N)_i) = \mathcal{N}\big(m_*^P + \Gamma_*^{P,O} (\Gamma_*^{O,O})^{-1} (y_*^O - m_*^O), \Gamma_*^{P,P} - \Gamma_*^{P,O} (\Gamma_*^{O,O})^{-1} \Gamma_*^{O,P}\big) \tag{2}$$

4 L'apprentissage

4.1 Etape E:

$$
\begin{aligned}
p(\mu_0^N|(y_i^N)_i,\Theta) &\propto \overbrace{p((y_i^N)_i|\mu_0^N,\sigma^2,(\theta_{\theta_i})_i))}^{\prod_{i=1}^{M}\mathcal{N}(\mu_0^N,\Psi_i^N)}\overbrace{p(\mu_0^N|\theta_0)}^{\mathcal{N}(0,K_{\theta_0})} \\
\ln p(\mu_0^N|(y_i^N)_i,\Theta) &= -\frac{1}{2}\mu_0^{N^T}K_{\theta_0}^{-1}\mu_0^N - \frac{1}{2}\sum_{i=1}^{M} -2y_i^{N^T}\Psi_i^{N^{-1}}\mu_0^N + \mu_0^{N^T}\Psi_i^{N^{-1}}\mu_0^N + C_1 \\
&= -\frac{1}{2}\mu_0^{N^T}(K_{\theta_0}^{-1} + \sum_{i=1}^{M}\Psi_i^{N^{-1}})\mu_0^N - \mu_0^N\sum_{i=1}^{M}\Psi_i^{N^{-1}}y_i^N + C_1
\end{aligned}
$$

Ainsi:

$$p(\mu_0^N|(y_i^N)_i,\Theta) = \mathcal{N}(m^N,\hat{K}^N)$$

Avec:

- $\hat{K}^N = \left((K_{\theta_0}^N)^{-1} + \sum\limits_{i=1}^{M}(\Psi_i^N)^{-1}\right)^{-1}$
- $m^N = \hat{K}^N\sum\limits_{i=1}^{M}(\Psi_i^N)^{-1}y_i^N$

5 Les preuves associées

Pour les démonstrations, on va user et abuser du Lemme ci dessous.

Lemma 1. Pour un espace $\mathcal{X}$ qui permet d'utiliser Fubini ,

$$\forall A,B \in \mathcal{X}, \quad \mathbb{E}_{p(A)}\left[g(A)\right] = \mathbb{E}_{p(B)}\left[\mathbb{E}_{p(A|B)}\left[g(A)\right]\right]$$

Proof.

$$
\begin{aligned}
\mathbb{E}_{p(A)}\left[g(A)\right] &= \int g(A)p(A)dA \\
&= \int g(A)\int p(A,B)dB\ dA \\
&= \int\int g(A)p(A|B)p(B)dB\ dA \\
&\underbrace{=}_{Fubini} \int (\int g(A)p(A|B)dA)\ p(B)dB \\
&= \mathbb{E}_{p(B)}\left[\mathbb{E}_{p(A|B)}\left[g(A)\right]\right]
\end{aligned}
$$

□

5.1 Preuve de (1)

On sait que la convolution de deux gaussiennes est elle aussi gaussienne, il suffit donc de calculer ses paramètres à l'aide du Lemme 1 avec $A = y_*^N|(y_i^N)_i$ et $B = \mu_0^N|(y_i^N)_i$

$$
\begin{aligned}
m_*^N &= \mathbb{E}_{p(y_*^N|(y_i^N)_i)}\left[y_*^N\right] \\
&= \mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}\left[\mathbb{E}_{p(y_*^N|(y_i^N)_i,\mu_0^N)}\left[y_*^N\right]\right] \\
&= \mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}\left[\mathbb{E}_{p(y_*^N|y_i^{t_i},\mu_0^{t_i^*})}\left[y_*^N\right]\right] \\
&= \mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}\left[\mu_0^N\right] \\
&= m^N
\end{aligned}
$$

On calcule ensuite le moment d'ordre 2:

$$
\begin{aligned}
\mathbb{E}_{p(y_*^N|(y_i^N)_i)}\left[y_*^N {y_*^N}^T\right] &= \mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}\left[\mathbb{E}_{p(y_*^N|(y_i^N)_i,\mu_0^N)}\left[y_*^N {y_*^N}^T\right]\right] \\
&= \mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}\left[\underbrace{\Psi_*^N}_{\mathbb{V}_{p(y_*^N|(y_i^N)_i)}[y_*^N]} + \underbrace{\mu_0^N {\mu_0^N}^T}_{\mathbb{E}_{p(y_*^N|(y_i^N)_i)}[y_*^N]\mathbb{E}_{p(y_*^N|(y_i^N)_i)}[{y_*^N}^T]}\right] \\
&= \Psi_*^N + \mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}\left[\mu_0^N {\mu_0^N}^T\right] \\
&= \Psi_*^N + \underbrace{\hat{K}^N}_{\mathbb{V}_{p(\mu_0^N|(y_i^N)_i)}(\mu_0^N)} + \underbrace{m^N {m^N}^T}_{\mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}[\mu_0^N]\mathbb{E}_{p(\mu_0^N|(y_i^N)_i)}[{\mu_0^N}^T]}
\end{aligned}
$$

Ainsi:

$$
\begin{aligned}
\Gamma_*^N &= \mathbb{V}_{p(y_*^N|(y_i^N)_i)}\left[y_*^N\right] \\
&= \mathbb{E}_{p(y_*^N|(y_i^N)_i)}\left[y_*^N {y_*^N}^T\right] - \mathbb{E}_{p(y_*^N|(y_i^N)_i)}\left[y_*^N\right]\mathbb{E}_{p(y_*^N|(y_i^N)_i)}\left[{y_*^N}^T\right] \\
&= \Psi_*^N + \hat{K}^N + m^N {m^N}^T - m^N {m^N}^T \\
&= \Psi_*^N + \hat{K}^N
\end{aligned}
$$

6 Model with clustering - variational EM

6.1 Notations

- M le nombre d'individus,
- N le nombre d'observation par individu,
- i, j les indices pour les individus,
- k, l les indices pour les observations,

On dispose d'un échantillon $\{(y_i^N, t_i^N)\}_{i=1,\dots,M}$, tel que:

- $t^N = (t_1, \dots, t_N)^T$ le vecteur des temps de l'individu i (communs à tous),
- $y_i^N = y_i(t^N) = (y_i(t_1), \dots, y_i(t_N))^T$ le vecteur des observations de l'individu i,
- $t^O \in \mathbb{R}^O,\ t^O \subset t^N$ le vecteur des temps d'observation d'un nouvel individu *,
- $t^P \in \mathbb{R}^P,\ t^P \subset t^N$ le vecteur des temps à prédire d'un nouvel individu *,
- $O\ + P = N$ et $t_i^O \cup t_i^P = t_i^N, \forall i$
- $y_i^O = y_i(t^O)$, $y_i^P = y_i(t^P)\ , \forall i$
- K_{θ_0} une noyau d'hyperparamètres θ_0

- $(\Sigma_{\theta_i})_i$ un ensemble de noyau de même forme et d'hyperparamètres respectifs $(\theta_i)_i$
- $\sigma^2 \in R$
- $\Theta = \{\theta_0, (\theta_i)_i, \sigma^2\}$ le vecteur des hyperparamètres du modèle

L'utilisation des indices avec les autres objets est cohérente avec celle ci-dessus.

6.2 Learning

6.2.1 E step

$$\log p\left(y|\pi,\theta,\gamma,\sigma^2\right) = \mathcal{L}\left(r(\mu,Z);\pi,\theta,\gamma,\sigma^2\right) + KL(r(\mu,Z)||p\left(\mu,Z)|y,\pi,\theta,\gamma,\sigma^2\right)$$

We assume that $r(\mu,Z) = r(\mu)r(Z)$. Considering this variational hypothesis, we know from *ref variational*, that the optimal functions are $\hat{r}(\mu)$, and $\hat{r}(Z)$ defined as such:

6.2.1.1 Calculation of $\hat{r}(\mu)$

$$\begin{aligned}
\log \hat{r}(\mu) &= \mathbb{E}_Z\left[\log p\left(y,Z,\mu|\pi,\theta,\gamma,\sigma^2\right)\right] + C_1 \\
&= \mathbb{E}_Z\left[\log p\left(y|Z,\mu,\pi,\gamma,\sigma^2\right)\right] + \log p(\mu|\theta) + C_2 \\
&= \sum_{i=1}^{M} \mathbb{E}_{Z_i}\left[\log p\left(y_i|Z_i,\mu,\gamma_i,\sigma_i^2\right)\right] + \sum_{q=1}^{Q} \log p(\mu_q|\theta_q) + C_2 \\
&= \sum_{i=1}^{M} \mathbb{E}_{Z_i}\left[\sum_{q=1}^{Q} Z_{iq} \log p\left(y_i|\mu_q,\gamma_i,\sigma_i^2\right)\right] + \sum_{q=1}^{Q} \log p(\mu_q|\theta_q) + C_2 \\
&= \sum_{q=1}^{Q}\left\{-\frac{1}{2}(\mu_q - m_q)^T K_{\theta_q}^{-1}(\mu_q - m_q) - \frac{1}{2}\sum_{i=1}^{M} \tau_{iq}\left(y_i - \mu_q\right)\Psi_{\gamma_i,\sigma_i^2}^{-1}\left(y_i - \mu_q\right)\right\} + C_3 \\
&= -\frac{1}{2}\mu_q^T\left(K_{\theta_q}^{-1} + \sum_{i=1}^{M}\tau_{iq}\Psi_{\gamma_i,\sigma_i^2}^{-1}\right)\mu_q + \mu_q^T\left(K_{\theta_q}^{-1}m_q + \sum_{i=1}^{M}\tau_{iq}\Psi_{\gamma_i,\sigma_i^2}^{-1}y_i\right) + C_4
\end{aligned}$$

We identify a gaussian likelihood and thus:

$$\begin{aligned}
\hat{r}(\mu) &= \prod_{q=1}^{Q} \hat{r}(\mu_q) \\
&= \prod_{q=1}^{Q} \mathcal{N}(\mu_q; \hat{\mu}_q, \hat{K}_q),
\end{aligned}$$

with:

- $\hat{K}_q = \left(K_{\theta_q}^{-1} + \sum_{i=1}^{M}\tau_{iq}\Psi_{\gamma_i,\sigma_i^2}^{-1}\right)^{-1}$,
- $\hat{\mu}_q = \hat{K}_q\left(K_{\theta_q}^{-1}m_q + \sum\limits_{i=1}^{M}\tau_{iq}\Psi_{\gamma_i,\sigma_i^2}^{-1}y_i\right)$.

6.2.1.2 Calculation of $\hat{r}(Z)$

$$
\begin{aligned}
\log \hat{r}(Z) &= \mathbb{E}_\mu \left[\log p\left(y, Z, \mu | \pi, \theta, \gamma, \sigma^2\right)\right] + C_1 \\
&= \mathbb{E}_\mu \left[\log p\left(y | Z, \mu, \gamma, \sigma^2\right)\right] + \log p(Z|\pi) + C_2 \\
&= \mathbb{E}_\mu \left[\sum_{i=1}^{M}\sum_{q=1}^{Q} Z_{iq} \log p\left(y_i | \mu_q, \gamma_i, \sigma_i^2\right)\right] + \sum_{i=1}^{M}\sum_{q=1}^{Q} Z_{iq} \log(\pi_q) + C_2 \\
&= \sum_{i=1}^{M}\sum_{q=1}^{Q} Z_{iq} \left\{\log(\pi_q) - \frac{1}{2}\left(\log|\Psi_{\gamma_i,\sigma_i^2}| + \mathbb{E}_\mu\left[\left(y_i - \mu_q\right)\Psi^{-1}_{\gamma_i,\sigma_i^2}\left(y_i - \mu_q\right)\right]\right)\right\} + C_3 \\
&= \sum_{i=1}^{M}\sum_{q=1}^{Q} Z_{iq} \left\{\log(\pi_q) - \frac{1}{2}\left(\log|\Psi_{\gamma_i,\sigma_i^2}| + \left(y_i - \hat{\mu}_q\right)\Psi^{-1}_{\gamma_i,\sigma_i^2}\left(y_i - \hat{\mu}_q\right) - Tr(\Psi^{-1}_{\gamma_i,\sigma_i^2}\hat{K}_q)\right)\right\} + C_3
\end{aligned}
$$

By identification, we found that $\hat{r}(Z) = \sum_{i=1}^{M} \hat{r}(Z_i)$ with:

- $\hat{r}(Z_i) = \mathcal{M}(1, \tau_{iq})$,
- $\log \tau_{iq} = \log(\pi_q) - \frac{1}{2}\left(\log|\Psi_{\gamma_i,\sigma_i^2}| + \left(y_i - \hat{\mu}_q\right)\Psi^{-1}_{\gamma_i,\sigma_i^2}\left(y_i - \hat{\mu}_q\right) - Tr(\Psi^{-1}_{\gamma_i,\sigma_i^2}\hat{K}_q)\right) + C_4$,
- $\tau_{iq} = \dfrac{\pi_q \mathcal{N}(y_i, \hat{\mu}_q, \Psi_{\gamma_i,\sigma_i^2}) \exp\left(-\frac{1}{2}Tr(\Psi^{-1}_{\gamma_i,\sigma_i^2}\hat{K}_q)\right)}{\sum_{l=1}^{Q} \pi_l \mathcal{N}(y_i, \hat{\mu}_l, \Psi_{\gamma_i,\sigma_i^2}) \exp\left(-\frac{1}{2}Tr(\Psi^{-1}_{\gamma_i,\sigma_i^2}\hat{K}_l)\right)}$.

6.2.2 M step

By explicit optimization, we get that:

$$
\hat{\pi}_q = \frac{1}{M}\sum_{i=1}^{M} \tau_{iq}, \quad \forall q = 1, \dots, Q.
$$

Proposition 1. *Let* Θ *be* $\{\theta, \gamma, \sigma^2\}$,

$$
\begin{aligned}
\hat{\Theta} = arg\max_{\Theta} -\frac{1}{2}\sum_{i=1}^{M}&\left\{log|\Psi_{\gamma_i,\sigma_i^2}| + y_i^T \Psi^{-1}_{\gamma_i,\sigma_i^2} y_i - 2y_i^T \Psi^{-1}_{\gamma_i,\sigma_i^2}\sum_{q=1}^{Q}\tau_{iq}\hat{\mu}_q + Tr(\Psi^{-1}_{\gamma_i,\sigma_i^2}\sum_{q=1}^{Q}\tau_{iq}\hat{\mu}_q^T\hat{\mu}_q) + Tr(\Psi^{-1}_{\gamma_i,\sigma_i^2}\sum_{q=1}^{Q}\tau_{iq}\hat{K}_q)\right\} \\
&+ \sum_{q=1}^{Q}\left\{log|K_{\theta_q}| + (\hat{\mu}_q - m_q)^T K_{\theta_q}^{-1}(\hat{\mu}_q - m_q) + Tr(K_{\theta_q}^{-1}\hat{K}_q)\right\}
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\mathcal{L}\left(r(\mu, Z); \pi, \theta, \gamma, \sigma^2\right) &= \mathbb{E}_{\mu,Z}\left[\log p\left(y, Z, \mu | \pi, \theta, \gamma, \sigma^2\right)\right] - \underbrace{\mathbb{E}_{\mu,Z}\left[\log \ r(\mu, Z)\right]}_{constant} \\
&= \mathbb{E}_{\mu,Z}\left[\log p\left(y | Z, \mu, \gamma, \sigma^2\right)\right] + \mathbb{E}_{\mu}\left[\log p\left(\mu | \theta\right)\right] + \mathbb{E}_{Z}\left[\log p\left(Z | \pi\right)\right] + C_1 \\
&= \sum_{i=1}^{M} \sum_{q=1}^{Q} \tau_{iq} \left\{ -\frac{1}{2} log |\Psi_{\gamma_i, \sigma_i^2}| - \frac{1}{2}(y_i - \hat{\mu}_q)^T \Psi_{\gamma_i, \sigma_i^2}^{-1} (y_i - \hat{\mu}_q) - \frac{1}{2} Tr(\Psi_{\gamma_i, \sigma_i^2}^{-1} \hat{K}_q) \right\} \\
&\quad + \sum_{q=1}^{Q} \left\{ -\frac{1}{2} log |K_{\theta_q}| - \frac{1}{2}(\hat{\mu}_q - m_q)^T K_{\theta_q}^{-1} (\hat{\mu}_q - m_q) - \frac{1}{2} Tr(K_{\theta_q}^{-1} \hat{K}_q) \right\} \\
&\quad + \sum_{i=1}^{M} \tau_{iq} \log(\pi_q) + C_1
\end{aligned}
$$

□

6.3 Mean processes update

6.4 Prediction